Data Governance & Ethics Framework

1. Data Sources & Provenance

- **Historical Data:** Sourced from the UCI Machine Learning Repository (via Kaggle). It represents sensor readings from an Italian city (2004-2005).
  - *Assumption:* We assume the sensor calibration from 2004 provides a valid baseline for relative pollution trends, despite technological shifts.
- **Live Data:** Sourced from the Open-Meteo Air Quality API.
  - *Assumption:* The API data is "True Concentration" (µg/m³) and has not been pre-scaled. We acknowledge a geographical proxy (Rome) is used to represent "Italian Air Quality."

2. Ethical Considerations & Privacy

- **Public Data:** Both datasets are public domain. No Personally Identifiable Information (PII) is processed.
- **Algorithmic Bias:** The BQML model is trained on 2004 data. There is a risk of **Concept Drift**—environmental regulations implemented since 2004 may have fundamentally changed the ratio of NO2 to CO. The model should be used as a *reference*, not for critical health safety alerts, without further calibration.
- **Transparency:** All data transformations (unit conversions from mg to µg) are documented in BigQuery Views to ensure auditability.

3. Failure Playbook (Operational Resilience)

- **Scenario A: External API Failure (HTTP 429/500)**
  - *Symptom:* Cloud Function logs show connection errors.
  - *Response:* The Cloud Function implements **exponential backoff**. If the API rate limits us (429), the function pauses execution. Pub/Sub acts as a buffer; if the producer fails, no "bad data" enters the pipe. The Scheduler simply retries in the next 15-minute window.
- **Scenario B: Data Schema Drift**
  - *Symptom:* Pub/Sub messages fail to write to BigQuery (Dead Letter Queue fills up).
  - *Response:* The BigQuery Subscription is configured to drop mismatched schemas to protect table integrity. Alerts should be configured on the subscription's `unacked_messages` metric.
- **Scenario C: Model Degradation**
  - *Symptom:* Forecast Accuracy KPI in Dashboard drops below 80%.
  - *Response:* Trigger a manual retraining job (`bq/sql/1_train_model.sql`) to incorporate more recent data if available.

4. Maintenance Schedule

- **Monthly:** Review BigQuery costs and delete temp tables.
- **Quarterly:** Validate Open-Meteo API documentation for endpoint deprecations.